FAIR Principles and the Immune Epitope Database

James A. Overton; Knocean Inc.
Randi Vita, Alessandro Sette, Bjoern Peters; La Jolla Institute
Christopher J. Mungall; Lawrence Berkeley National Laboratory

2017-10-25

The Immune Epitope Database

The IEDB by the Numbers

A publicly available database of experiments demonstrating recognition of immune epitopes by adaptive immune receptors.

- established 2004 by NIAID
- based at the La Jolla Institute of Allergy and Immunology
- 18,804 journal articles + direct submissions
- 1,509,095 assays (individual experiments)
- up to 400 fields per assay
- 385,446 epitopes

Epitope	
Epitope ID	717
Chemical Type	Linear peptide
Linear Sequence	ADLMGYIPLVGAPLGGAARA
Starting Position	131
Ending Position	150
Source Molecule Name	polyprotein
Source Accession	BAA03375.1
Source Organism ID	31647
Source Organism	Hepatitis C virus subtype 1b

Reference	
Article Authors	Angus J MacDonald; Margaret Duffy; Miriam T Brady; Susan McKiernan; William Hall; John Hegarty; Michael Curry; Kingston H G Mills
Article Title	CD4 T helper type 1 and regulatory T cells induced against the epitopes on the core protein in hepatitis C virus-infected persons.

Reference Detail	
Reference ID	1001817
Abstract	The factors that determine persistence or clearance of hepatitis C virus (HCV) infection are poorly understood. The CD4 T cell responses to the HCV core protein were examined in a cohort of women infected with a single genotype of HCV. CD4 T cells from HCV-infected patients secreted interferon (IFN)-gamma in response to peptides from 4 immunodominant regions of the core protein, and these responses were strong er in persistently infected women. Interleukin (IL)-10 was also produced by CD4 T cells from HCV-infected subjects in response to the same core peptides. Furthermore, HCV core-specific CD4 T cell clones secreted either IFN-gamma or IL-10 but not IL-4. These findings demonstrate that T helper type 1 and regulatory T cells are induced against the same epitopes on the core protein during HCV infection.
Affiliations	Immune Regulation Research Group, Department of Biochemistry, Trinity College Dublin, Ireland.
Date	2002
Reference Type	Literature
PubMed ID	11920289
Journal	J Infect Dis
Journal Volume	185
Article Pages	720-7
Journal ISSN	1537-6613
Article Chemical List	Epitopes, T-Lymphocyte;HLA-DR Antigens;HLA-DRB1 Chains;Interleukin-5;Viral Core Proteins;nucleocapsid protein, Hepatitis C virus;Interleukin-10;Interferon-gamma
Article MeSH List	Adult; CD4-Positive T-Lymphocytes(immunology); Epitopes, T-Lymphocyte(immunology); Female; HLA-DR Antigens(genetics); HLA-DRB1 Chains; Hepatitis C(immunology); Humans; Interferon-gamma(biosynthesis); Interleukin-10(biosynthesis); Interleukin-5(biosynthesis); Middle Aged; Th1 Cells(immunology); Viral Core Proteins(chemistry; immunology)

Epitope	
Epitope ID	717
Chemical Type	Linear peptide
Linear Sequence	ADLMGYIPLVGAPLGGAARA
Starting Position	131
Ending Position	150
Source Molecule Name	polyprotein
Source Accession	BAA03375.1
Source Organism ID	31647
Source Organism	Hepatitis C virus subtype 1b

Epitope Reference Details	
Epitope Structure Defines	Epitope containing region/antigenic site
Epitope Evidence Code	Not determined
Epitope Name	Core 131-150
Reference Starting Position	131
Reference Ending Position	150
Location of Data in Reference	Figure 1

Immunization	
Host Organism ID	9606
Host Organism	Homo sapiens

Host Details	
Sex	F
Age	41-55
MHC Types present	HLA-DRB1*07:01;HLA-DQB1*05:01;HLA-DRB1*01:01;HLA-DQB1*02:01

1st In Vivo Process	
In Vivo Process Type	Occurrence of infectious disease
Disease Code	171
Disease State	hepatitis C
Disease Stage	Chronic;

1st Immunogen	
Epitope Relation	Source Organism
Object Type	Organism
Organism ID	31647
Organism	Hepatitis C virus subtype 1b

Immunogen Details	
Immunogen Evidence Code	Not determined
Immunogen Reference Name	Hepatitis C virus subtype 1b

Immunization Comments	
Immunization Comments	PBMC were harvested from a group of Irish women infected with HCV 1b through the administration of contaminated anti-D Ig.

T Cell Assay	
Qualitative Measurement	Positive
Method/Technique	ELISA
Measurement of	IFNg release

Effector Cells	
Effector Cell Tissue Type	Blood
Effector Cell Type	PBMC
Effector Cell Culture Conditions	Direct Ex Vivo

Antigen Presenting Cells	
Cell Tissue Type	Blood
Cell Type	PBMC
Cell Culture Conditions	Direct Ex Vivo

Antigen	
Epitope Relation	Epitope
Chemical Type	Linear peptide
Linear Sequence	ADLMGYIPLVGAPLGGAARA
Starting Position	131
Ending Position	150
Source Molecule Name	polyprotein
Source Accession	BAA03375.1
Source Organism ID	31647
Source Organism	Hepatitis C virus subtype 1b

Antigen Details	
Antigen Evidence Code	Exact match to reference information
Antigen Reference Name	Core 131-150

Assay Reference Details	
Assay Comments by IEDB Curator	PBMC from HCV infected patients produced IFNg in response to the epitope. Chronically infected patients (HCV PCR+) responded more often than patients who cleared the virus (HCV RNA-).
Location of Assay Data in Reference	Figure 2 and table 1

FAIR is FAIR

https://www.force11.org/group/fairgroup/fairprinciples

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

Summary: FAIR and Square

F1	Persistent IDs
F2	Rich Metadata
F3	Metadata Links to Data
F4	Metadata Indexed
A1	Open Protocols
A2	Metadata Remain Accessible
I1	Open Language
I2	Open Vocabularies
I3	Metadata Links to Metadata
R1	Rich Metadata
R1.1	License
R1.2	Provenance
R1.3	Community

Discussion: A FAIR Shake

Data and Metadata

The distinction between data and metadata is not always clear or helpful.

For the current purposes, the IEDB is a repository of metadata.

Principles and Practises

FAIR principles tell us **what** to do, and sometimes **why** to do it.

They do not tell us **how** to do it.

Shared principles are not enough for interoperability.

FAIR and OBO

The Open Biomedical Ontologies (OBO) community has a set of principles, much in the same spirit as FAIR.

More importantly, OBO also has shared best practises, common infrastructure, open source tooling.

We see an important role for OBO to articulate shared best practises that make the FAIR principles concrete and effective.

My Opinions

	Principle	Practise
F1	Persistent IDs	IRIs and a PURL system
F2	Rich Metadata	RDF(S), OWL, DC, PROV, OBO metadata
A1	Open Protocols	HTTP(S), RDF/OWL in XML, JSON-LD
I1	Language	RDF/OWL
I2	Vocabularies	OBO community
R1.1	License	Creative Commons: CC-BY or CC0
R1.2	Provenance	PROV
R1.3	Community	OBO community

The IEDB: FAIRly Good

F1. Persistent IDs

The IEDB has persistent URLs for assay, reference, epitope:

- http://www.iedb.org/assay/1288922
- http://www.iedb.org/reference/1001817
- http://www.iedb.org/epitope/717

F2. Rich Metadata

Providing rich metadata about experiments is the heart of the IEDB mission.

We can do better!

- use standardized predicates
- share richer modelling
- RDF/OWL representation

F3. Metadata Links to Data

The IEDB links to journal articles by PubMed ID.

We also capture more specific location within the document (as free text): figure X, table Y.

We can do better!

- standardized, formalized location data
- link out to PubMed Central structured representations

F4. Metadata Indexed

The IEDB is itself an indexed, searchable resource.

We have also submitted our data to:

- Biosharing
- bioCADDIE
- Wikidata

A1. Open Protocols

The IEDB is:

- accessible by HTTP, no authentication required
- published as HTML pages and CSV tables

We can do better!

- CSV with IRIs and labels
- RDF/OWL representation as JSON-LD or RDFa

Google Structured Data Testing Tool

http://8.37.117.76/epitope/123885

NEW TEST

```
epitopes. Includes more than 95% of all
published infectious disease, allergy,
autoimmune, and transplant epitope data.">
<meta name="twitter:image"
content="http://www.iedb.org/images/IEDB.png">
<meta name="twitter:card" content="summary">
<!-- JSON-LD tag -->
<script type="application/ld+json">
{
    "http://purl.org/dc/terms/license":
"https://creativecommons.org/licenses/by/4.0/"
,
"http://www.w3.org/ns/prov#wasGeneratedBy":
"https://scicrunch.org/browse/resources/SCR_00
6604"
}
</script>
```

Unspecified Type

All (1)

Unspecified Type

0 ERRORS 0 WARNINGS

@type	*Unspecified Type*
http://purl.org/dc/terms/license	https://creativecommons.org/licenses/by/4.0/
http://www.w3.org/ns/prov#wasGeneratedBy	https://scicrunch.org/browse/resources/SCR_006604

A2. Metadata Remain Accessible

Published journal articles remain accessible, even if retracted.

The IEDB metadata records will remain accessible in any case.

I1. Knowledge Representation Language

The IEDB uses SQL and publishes in HTML and CSV.

We can do better!

- CSV exports with IRIs and labels
- RDF/OWL representation in JSON-LD or RDFa

I2. FAIR Vocabularies: Reuse

The IEDB links to a wide range of resources:

- UniProt
- GenBank
- NCBI Taxonomy
- PDB
- IMGT
- many OBO ontologies
- more. . .

I2. FAIR Vocabularies: Contributions

The IEDB contributes whenever possible:

- Ontology for Biomedical Investigations (>300 terms)
- Disease Ontology (>200 terms)
- Chemical Entities of Biological Interest (>2400 terms)

I2. FAIR Vocabularies: Development

The IEDB has also developed the MHC Restriction Ontology, following OBO principles:

http://purl.obolibrary.org/obo/MRO

I2. FAIR Vocabularies: Public

Sometimes there is no good home for a term that we need: mouse/rat strains, post-translational modifications of proteins, etc.

We put these terms in an application ontology ONTIE, then migrate to better homes when available.

ONTIE is now public with PURLs and RDF/OWL representations in multiple formats:

https://ontology.iedb.org/ontology/ONTIE_0002032

ONTIE Terms

ONTIE:0002032 Polymerase basic protein 2 (Influenza A virus)

https://ontology.iedb.org/ontology/ONTIE_0002032

- type: owl:Class
- label: Polymerase basic protein 2 (Influenza A virus)
- alternative term: PB2
- alternative term: RNA-directed RNA polymerase subunit P3
- IEDB alternative term: Polymerase basic protein 2
- template: protein class
- ONTIE domain: protein
- protein label: Polymerase basic protein 2
- protein taxon: Influenza A virus

Other formats: Turtle (ttl), JSON-LD (json), TSV (tsv).

I2. FAIR Vocabularies: Improvements

We can do better!

We want to do a better job of explaining to users how vocabularies are used.

One example: our assays often compose a method with a GO process.

I3. Metadata Link to Metadata

The IEDB links to many resources (I2), which link to other resources, which link to other resources, . . .

R1.1. License

The IEDB was failing this principle.

The IEDB's content is available under the Creative Commons Attribution 4.0 International license. We are making this clear with human- and machine-readable annotations.

R1.2. Provenance

The IEDB uses PubMed IDs to link to the original article.

We are using PROV and adding metadata about our generation of the metadata from an original source.

R1.3. Community

We are careful to use standards that are appropriate for the immunology community.

Summary: Playing FAIR

- F3: Standardize identification of journal parts (Figures / Tables)
- F4: Add IEDB metadata to Biosharing, Biocaddie and Wikidata
- A1: Provide machine actionable representation of IEDB assay level data
- I1: Represent the IEDB data in RDF/OWL
- I2: Make all links to external vocabularies explicit
- I2: Make all internal vocabularies public via ONTIE and link to them
- R1.1: Include licensing information with the IEDB records
- R1.2: Include provenance information regarding IEDB curation

More than FAIR

More than FAIR

FAIR is not enough.

Interoperability requires shared best practises, compatible implementations.

Full interoperability also requires shared modelling patterns – we must represent similar things in similar ways.

We see a role for OBO here, and have started coordinated modelling of immunology datasets from the IEDB, BRCs, and ImmPort.

Acknowledgements

Funding is provided by The National Institutes of Health

HHSN272201200010C

FAIR Thee Well